Broadband Wireless Data and the "User Experience"

Dr. Jay E. Padgett
Applied Research
Telcordia Technologies, Inc.

September 21, 2004

OVERVIEW

Problem Summary and Objectives

There are *N* users who share a wireless channel of capacity R on a time-division basis. While the *average* rate is $R_{avg} = R/N$, this is not necessarily the apparent or effective rate experienced by the user. If the user is engaged in activities that demand bursts of data followed by inactive periods, such as downloading and reading web pages, then the channel rate as perceived by the user depends on how long it takes to load the web page following the user's request. The goal of this paper is to understand how this effective rate R_{eff} depends on *R* and *N*, and the parameters of the data traffic. The purpose here is not to provide a rigorous or exhaustive analysis, but rather to illustrate, with simple approximations and an elementary simulation, the relationship among channel rate, number of active users, idle time, data block size, and the effective rate.

Assumed Operating Scenario and Effective Rate Definition

To illustrate the basic principles, a simple web-browsing scenario is assumed here as suggested in [1]. With this model, users are downloading and reading web pages. Each page requires L kb of downloaded data, and the user absorption time for each page is t_a. Following [1] it is assumed here initially that $L = 800\,\text{kb}$ and $t_a = 30\,\text{sec}$. In other words, each user will download an 800-kb page, study it for 30 seconds, and then request another page download. Other data [2] suggest that 50-60 seconds is a more realistic value for capacity calculations, so results are also shown for $t_a = 50$ seconds.

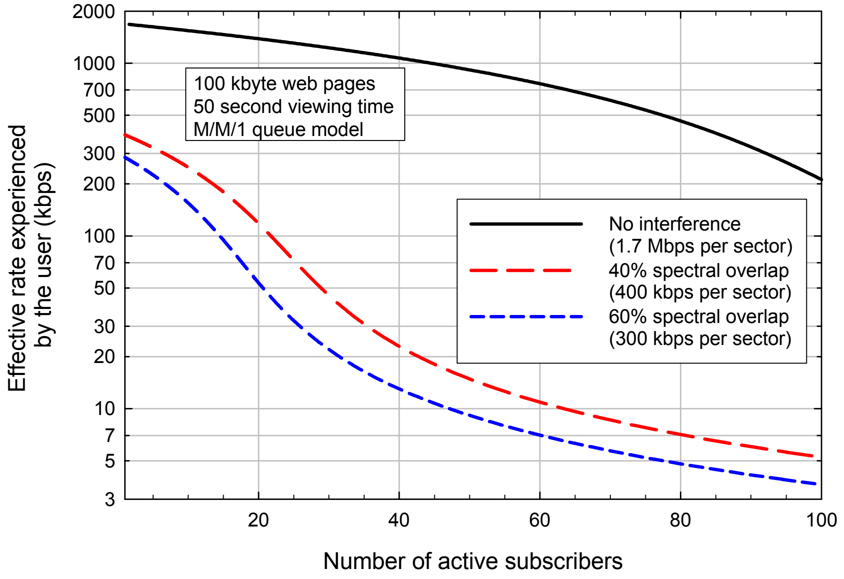
The service time required to actually load the web page is $t_{svc} = L/R$. It is assumed that requests are handled on a first-come, first-served basis and requests that cannot be served immediately join a queue. The total time required for the download as experienced by the user is then the time spent waiting in the queue, denoted t_{wait}, plus the actual service time: $t_q = t_{wait} + t_{svc}$. It is assumed here that the effective data rate experienced by the user is then $R_{eff} = L/t_q$.

Conclusions

Based on this initial analysis, it is evident that:

- There is a pronounced threshold, or saturation, effect. As the number of users increases, there is a break point below which there is little delay and above which delay increases linearly with the number of additional users, and the effective rate to the user decreases accordingly. A simple linear approximation for the waiting delay above the break point agrees well with the simulation results, for *N* greater than 20-30% above the break point.
- The effective rate as experienced by the user will always exceed the average rate (available rate divided by the number users), by an amount that depends on the number of users relative to the saturation point.
- If the channel rate is changed by some factor, the effective rate to the user changes by a larger factor.

The immediate application of this model is to assess the effective rate to the user provided by a broadband wireless air-to-ground (ATG) service under various conditions. Proposals have been made [6] to share the ATG bands among multiple providers using spectral overlap. This will cause interference among the providers and reduce available throughput. Based on simulations discussed elsewhere [7], the forward link throughput is 1.7 Mpbs without spectrum sharing, 400 kbps with 40% spectral overlap, and 300 kbps with 60% spectral overlap. The graph below shows how these three values of total throughput per sector translate to the rate experienced by the user. For these curves, the model based on the M/M/1 queue was used, which gives the most conservative results. However, as will be seen, there is little difference between these results and those based on the M/D/1 queue, or the simulation results.



ANALYSIS AND SIMULATION DETAILS

Queueing Model Approximations

The goal is to develop a relationship that shows R_{eff} as a function of *N*, given *R, L,* and t_a. An obvious first step is to model the situation as a single-server queue. Kleinrock [4] gives expressions for the average number of messages in a single-server queue for Poisson arrivals, and both exponentially-distributed and fixed-length service times. These are denoted, respectively, M/M/1 and M/D/1 queues. If λ is the average request arrival rate, and τ is the average service time, then the traffic intensity is $\rho = \lambda\tau$, and the average number of queued messages for the two cases are:

$$\bar{n}_q = \frac{\rho}{1-\rho} \qquad \text{M/M/1} \tag{1}$$

$$\bar{n}_q = \frac{\rho}{1-\rho} - \frac{\rho^2}{2(1-\rho)} \quad \text{M/D/1} \tag{2}$$

These represent the average number of messages in queue (including those being served) at an arbitrary instant in time. As noted in [4], the average number waiting but not being served is $\bar{n}_q - \rho$. As also explained in [4], Little's theorem gives the average time spent in queue as

$$\bar{t}_q = \frac{\bar{n}_q}{\lambda}. \tag{3}$$

Hence, the average waiting time (the time spent in queue but not being served) is $\bar{t}_{wait} = (\bar{n}_q - \rho)/\lambda = (\bar{n}_q/\lambda) - \tau$.

It is sometimes useful to normalize to the mean service time τ:

$$\frac{\bar{t}_q}{\tau} = \frac{\bar{n}_q}{\rho} \qquad\qquad \frac{\bar{t}_{wait}}{\tau} = \frac{\bar{n}_q}{\rho} - 1 \tag{4}$$

In the problem of interest here, there are *N* active users. If on average, $\bar{n}_q$ of them are in the queue, then the average arrival rate is

$$\lambda = \frac{N - \overline{n}_q}{t_a}. \tag{5}$$

Hence,

$$\rho = \lambda\tau = \left(N - \overline{n}_q\right)\frac{\tau}{t_a}, \tag{6}$$

so

$$N = \rho\frac{t_a}{\tau} + \overline{n}_q \ . \tag{7}$$

The average queue time can be shown as a function of *N* by specifying ρ, computing *N* according to (7), and then computing t_q/τ according to (4) as:[1]

$$\frac{\bar{t}_q}{\tau} = \frac{\overline{n}_q}{\rho} = \begin{cases} \dfrac{1}{1-\rho} & \text{M/M/1} \\ \dfrac{1}{1-\rho} - \dfrac{\rho}{2(1-\rho)} & \text{M/D/1} \end{cases}. \tag{8}$$

The "effective" rate as experienced by the user can be expressed as

$$R_{eff} = \frac{L}{\bar{t}_q} = \frac{\tau R}{\bar{t}_q} = \frac{R}{\overline{n}_q/\rho}. \tag{9}$$

The average rate is simply

$$R_{avg} = \frac{R}{N}, \tag{10}$$

and the "oversubscription" or "overbooking" factor can be defined as

[1] Strictly speaking, this analysis could be performed more rigorously by modeling the input stream as quasirandom rather than Poisson, and developing the queue state probabilities (see [5], chapter 3). However, the approach used here is simpler and is adequate for the present purpose.

$$f_{ovbk} = \frac{R_{eff}}{R_{avg}}. \tag{11}$$

Simulation

Simulating this situation is straightforward, especially if the service time is assumed constant as is the case here. Time is divided into intervals equal to the service time: $\Delta t = t_{svc}$, and then subdivided into 100 subintervals $\delta t = \Delta t / 100$, to provide sufficiently fine granularity for simulating the request arrival process. In each of these subintervals, the probability of an arriving service request from a single idle customer is

$$p_1 = \frac{\delta t}{t_a}. \tag{12}$$

If there are n_{idle} idle customers, then the probability that none of them will request service is:

$$p_0 = (1 - p_1)^{n_{idle}}, \tag{13}$$

and the probability of a request in the interval δt is $p_{req} = 1 - p_0$. With $\lambda = n_{idle} / t_a$, note that, consistent with the Poisson arrival model, $p_{req} \cong n_{idle} p_1 = \lambda \cdot \delta t$ for $\lambda \cdot \delta t << 1$. Since $\lambda_{\max} = 1/t_{svc}$, which occurs when $\rho = 1$ (the server is fully occupied), $\lambda \cdot \delta t \le 0.01$ in this case, so the time resolution is sufficiently fine to simulate Poisson arrivals.

In the simulation, a random number *u* that is uniformly-distributed on (0, 1) is generated, and if $u < p_{req}$, a message is added to the queue and n_q (the number of message currently in queue) is incremented. This arriving-request simulation procedure is executed 100 times per interval Δt. Following that, if there is a message in the first position of the queue (the service position), its total queue time is added to the running sum queue time (used for computing average queue time), the service counter is incremented, and all other messages in the queue are advanced one position and their queue times are increased by one service time interval.

What is described above is a single pass through the loop, and typically, 10,000 or more such iterations are used per value of *N*. After all iterations are complete, the running queue time sum is divided by the service counter to give the average queue time, which can be compared with that from the approximation described above, and R_{eff} can be computed according to (9). The queue array and all counters are then cleared and the simulation is repeated for the next value of *N*.

Figure 1 and Figure 2 show the average queue time (relative to the service time) for $R = 2$ Mbps and 1 Mbps, respectively, from the analysis described above and from the simulation. As can be seen, agreement between the analysis and simulation is good, especially for the analysis based on the M/D/1 queue (this is not surprising since fixed-length messages were used in the simulation). A pronounced break point is evident on all curves and can be explained intuitively as follows. If a perfect scheduling algorithm allowed users to make requests only when the previous transmission is complete, then the server would be fully occupied if $N = t_a / t_{svc}$. If *N* exceeds this value, then some users must wait before being served. The larger *N* becomes, the larger the queue becomes, because only t_a / t_{svc} users can be served every t_a seconds. As can be seen, the break point occurs in both cases is:

$$N_{break} = \frac{t_a}{t_{svc}} . \tag{14}$$

The slope of the curve beyond the break point can be derived using (7), with $\rho = 1$, since the server is always busy after the break point, and with $\tau = t_{svc}$:

$$N = \frac{t_a}{t_{svc}} + \overline{n}_q = N_{break} + \overline{n}_q \tag{15}$$

Thus, $\overline{n}_q = N - N_{break}$, and from (3), $\bar{t}_q = \overline{n}_q / \lambda$. Since $\lambda = \rho / t_{svc}$ and $\rho = 1$ at this point in the curve, $\overline{n}_q = \bar{t}_q / t_{svc}$, and a rough piecewise linear approximation can be written as:

$$\frac{\bar{t}_q}{t_{svc}} \cong \begin{cases} 1 & N \le N_{break} \\ N - N_{break} & N > N_{break} \end{cases} \tag{16}$$

As can be seen from Figure 1 and Figure 2, this gives a good fit to the simulation results except at the bend in the curve near the break point.

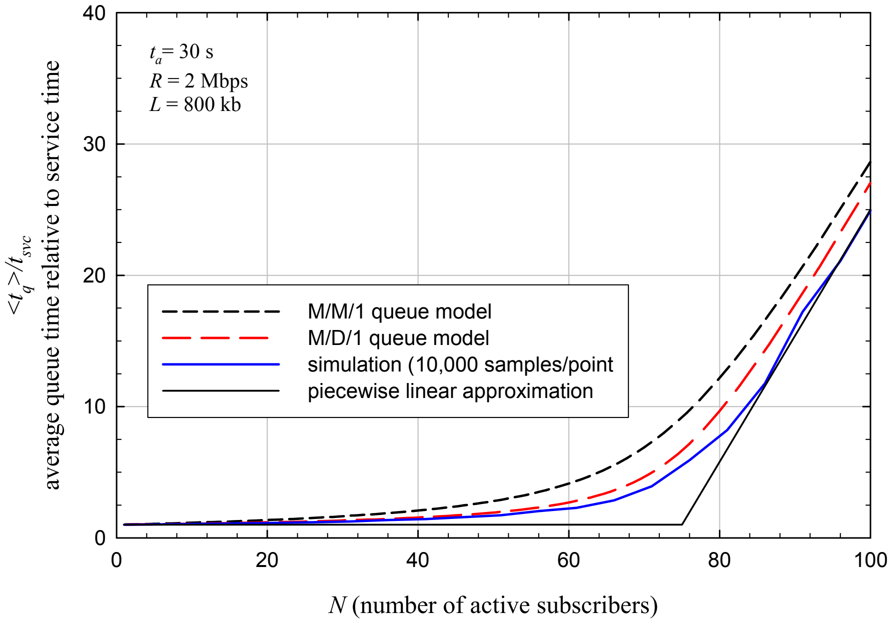


Figure 1: *Average queue time from the approximate model and from simulation for a total channel rate of 2 Mbps.*

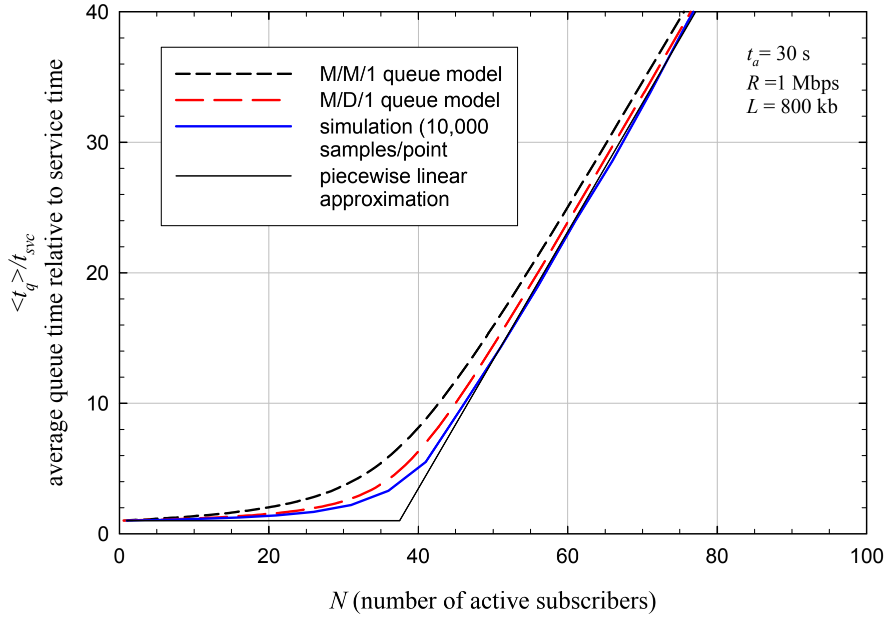


Figure 2: *Average queue time for a total channel rate of 1 Mbps.*

Figure 3 shows R_{eff} vs. the number of active users from both analysis and simulation, for $R = 1$ Mbps and $R = 2$ Mbps. The dashed curves show R_{eff} computed according to (9) where the average queue time $\bar{t}_q$ is computed in the simulation and $R_{eff} = L/\bar{t}_q$ is calculated at the end.

It is also possible to calculate R_{eff} in a different way, by computing L/t_q for each iteration, and accumulating a running sum, giving the average

$$R'_{eff} = \left\langle \frac{L}{t_q} \right\rangle = L\left\langle t_q^{-1} \right\rangle \tag{17}$$

effectively computing the harmonic mean $\left\langle t_q^{-1} \right\rangle^{-1}$ of the queue time rather than the mean $\left\langle t_q \right\rangle$. This is also shown in Figure 3 (dotted lines). While R'_{eff} is somewhat higher than R_{eff}, it is still subject to the same break-point behavior.

Finally, Figure 4 shows the so-called “oversubscription factor” or “overbooking factor”, which is simply the ratio R_{eff}/R_{avg} .

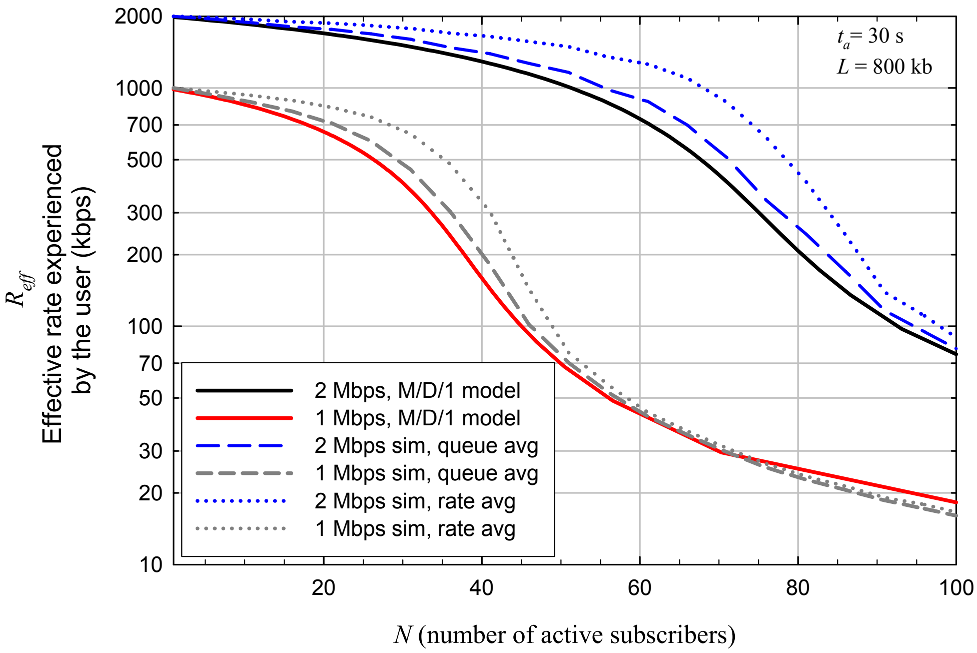


Figure 3: *Effective data rate to user from analysis and simulation.*

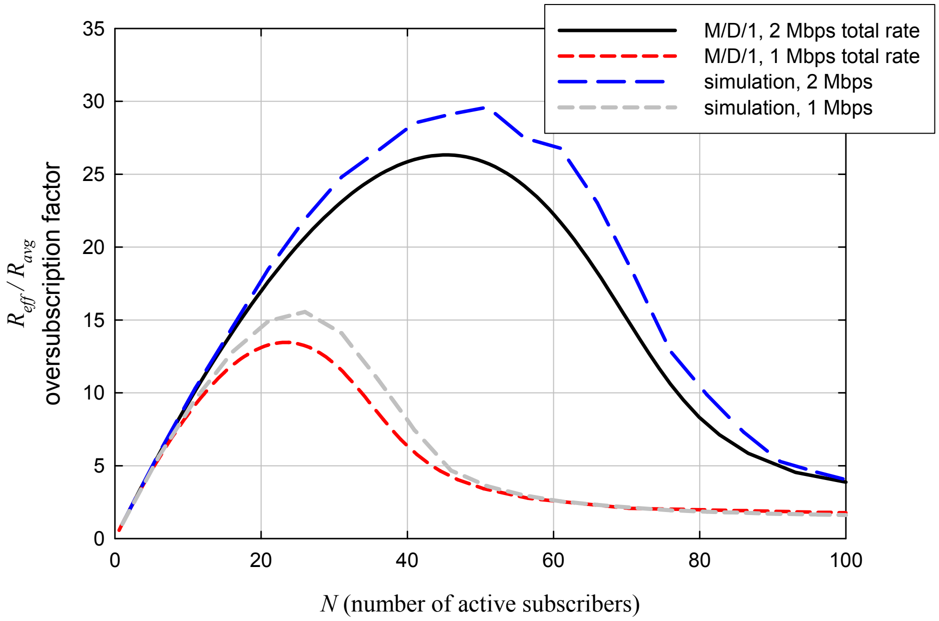


Figure 4: *Oversubscription or overbooking factor as a function of number of active users.*

Sensitivity Analysis

It is easy to understand the sensitivity to parameter variations from the break point analysis. The break point can be written as:

$$N_{break} = \frac{t_a R}{L}, \tag{18}$$

and from (9), $R_{eff} = R\, t_{svc} / \bar{t}_q$. From (16), $\bar{t}_q / t_{svc} \cong N - N_{break}$ for $N > N_{break}$. Therefore, the break point approximation gives

$$\frac{R_{eff}}{R} \cong \frac{1}{N - t_a R/L}, \qquad N > t_a R/L. \tag{19}$$

Clearly, it is $t_a R/L$ that controls R_{eff} relative to the total rate R. Figure 5 shows an example in which $L = 1600$ kb (double the previous case), and other parameters are unchanged. It can be seen that the break point has shifted left by a factor of 2. Also shown, for $R = 1$ Mbps, is the break-point approximation of (19) for $N \geq 1.5 N_{break}$. As can be seen, its agreement with the simulation results is excellent.

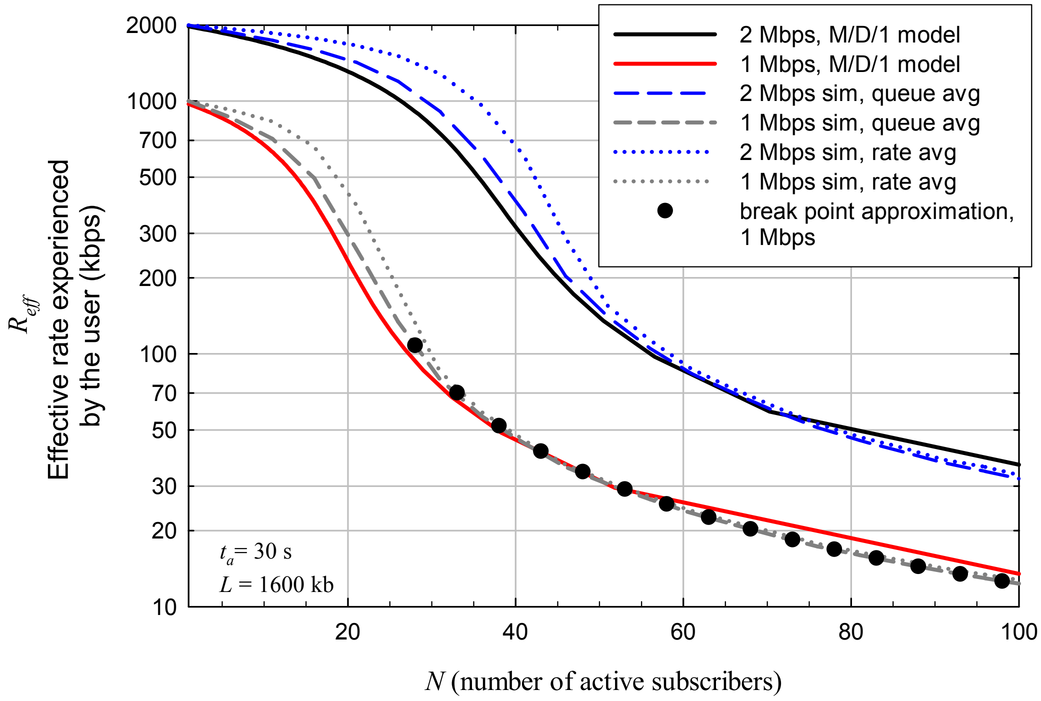


Figure 5: *Same as Figure 3, but with a block length of 1600 kb rather than 800 kb.*

Note that R_{eff} can be written as

$$\begin{aligned} R_{eff} &\cong \frac{1}{N/R - t_a/L}, \qquad N/R > t_a/L \\ &= \frac{1}{1/R_{avg} - t_a/L}, \qquad R_{avg} < L/t_a \end{aligned} \tag{20}$$

which shows that the effective rate always exceeds the average rate. In fact, the overbooking factor can be written as

$$\begin{aligned} f_{ovbk} = \frac{R_{eff}}{R_{avg}} &\cong \frac{1}{1 - Rt_a/NL} \\ &= \frac{N/N_{break}}{N/N_{break} - 1} \end{aligned} \qquad N > Rt_a/L\,. \tag{21}$$

This allows the value of *N* corresponding to a particular overbooking factor to be simply calculated:

$$N = N_{break} \cdot \frac{f_{ovbk}}{f_{ovbk} - 1}, \qquad N > N_{break}. \tag{22}$$

Figure 6 shows the approximation of (21) for the 1 Mpbs case for $N \geq 1.2 N_{break}$.

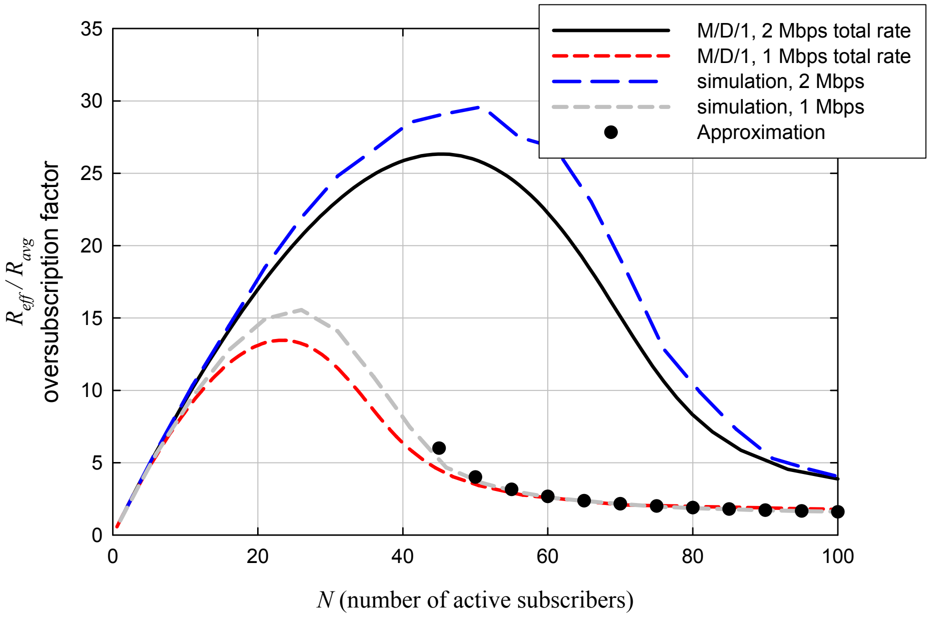


Figure 6: *Same as Figure 4, but with the approximation for R = 1 Mbps.*

Application to Broadband Air-to-Ground Communications

In [6], it was proposed that two air-to-ground (ATG) providers could share spectrum (with partial spectral overlap) by using "cross-duplexing" or "reverse banding" whereby the aircraft of one system transmits in the same band as the base station of the other system. This arrangement creates the potential for aircraft-to-aircraft interference, which can affect the reception on the forward (base to aircraft) link. This interference effect was quantified in [7], and Figure 7 shows the mean forward link throughput per sector for one system as a function of the aircraft deployment of the other system. The baseline case (a single system operating in exclusive spectrum) has a throughput of 1.7 Mbps per sector. When the second system has captured half the addressable market, the total throughput is reduced to about 400 kbps, if the maximum aircraft transmit power is 43 dBm (representing a high data rate on the reverse link) and the spectral overlap is 40%. If the spectral overlap is a more realistic 60% (to account for two 1.25 MHz blocks with a 125-kHz guard band at each edge of the 2 MHz ATG block), then the total throughput is about 300 kbps (not shown, but determined using the same simulation described in [7] which produced Figure 7).

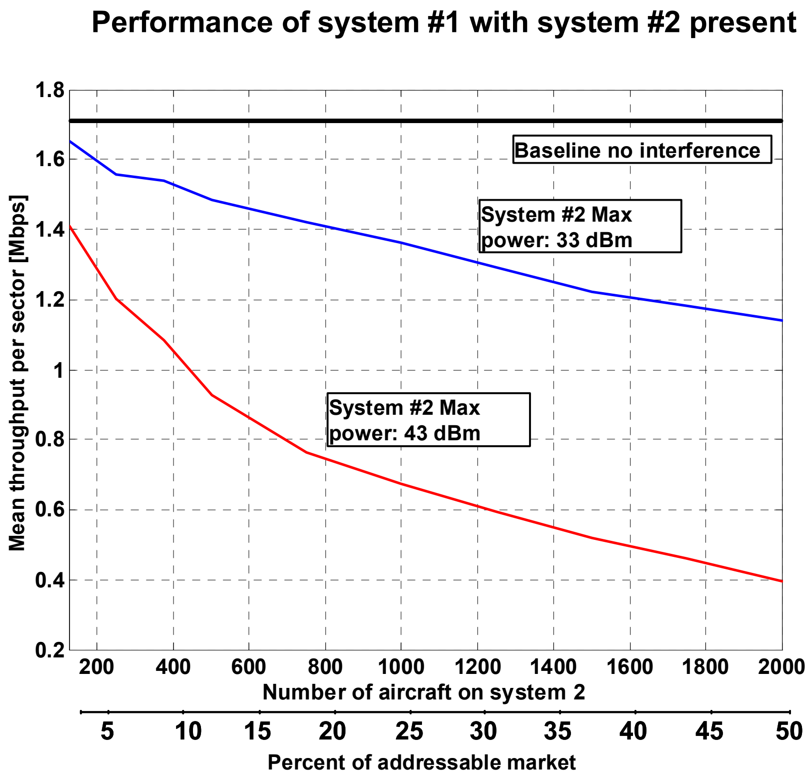


Figure 7: *Effect of air-to-air cross-duplex interference on mean forward link throughput with 40% spectral overlap (reproduced from Figure 40 of [7]).*

What is of interest here is the effect of the change in the throughput per sector on the effective rate as experienced by the user. This is straightforward to determine using the analysis developed here. However, the absorption time parameter used here will be 50 rather than 30 seconds, which appears more realistic based on [2]. From [3], the 800-kb page size used here may be slightly higher than typical, but as noted in [3], there is a tendency for average page size to increase over the years, so the 800 kb (100 kB) page size will be retained for the calculations.

Figure 8 shows R_{eff} vs. *N*. for $R = 1.7$ Mbps and $R = 400$ kbps (40% spectral overlap), and Figure 9 shows the curves for $R = 1.7$ Mbps and $R = 300$ kbps (60% spectral overlap). Results for the M/M/1 queue model are also shown, representing the effect of variable data block size. As would be expected, from the delay curves, the M/M/1 model results are the most conservative. Finally, Figure 10 shows the M/M/1 results for all three rates.

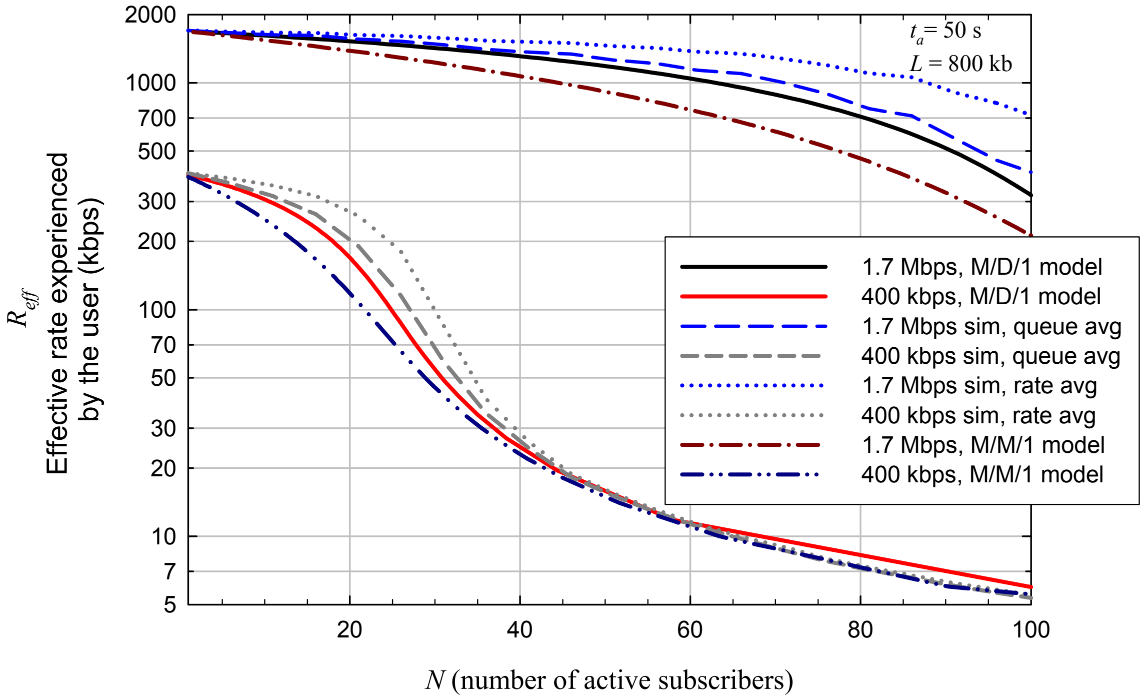


Figure 8: *Effective rate vs. number of users for 1.7 Mbps and 400 kbps (40% spectral overlap).*

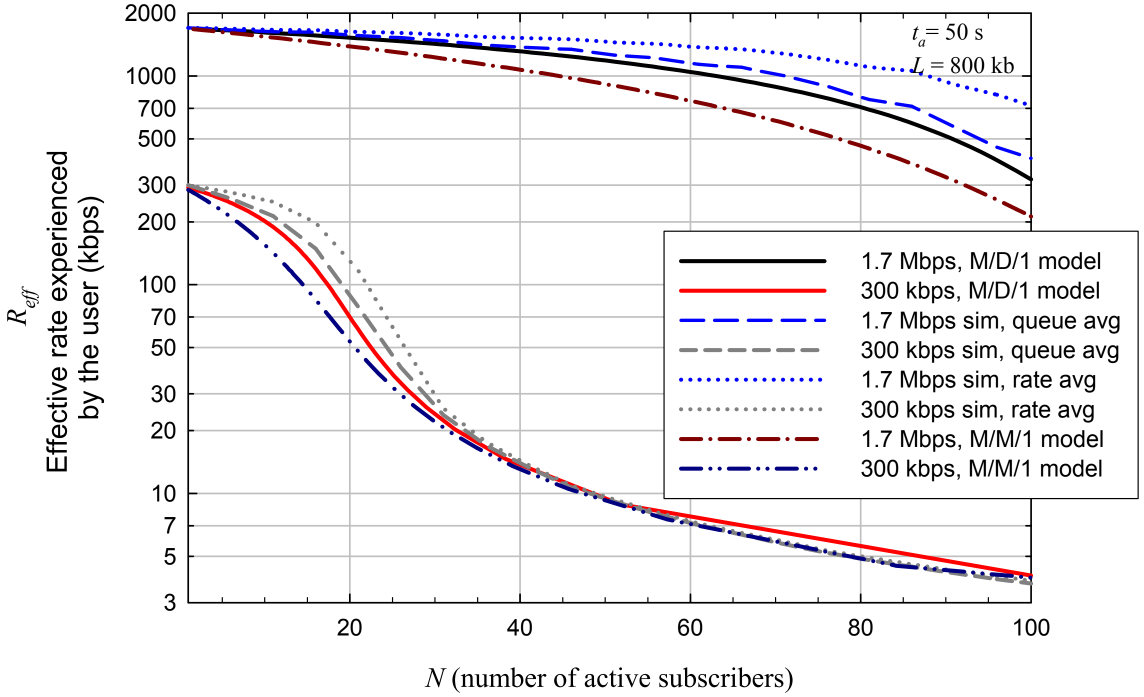


Figure 9: *Effective rate for 1.7 Mbps and 300 kbps (60% spectral overlap).*

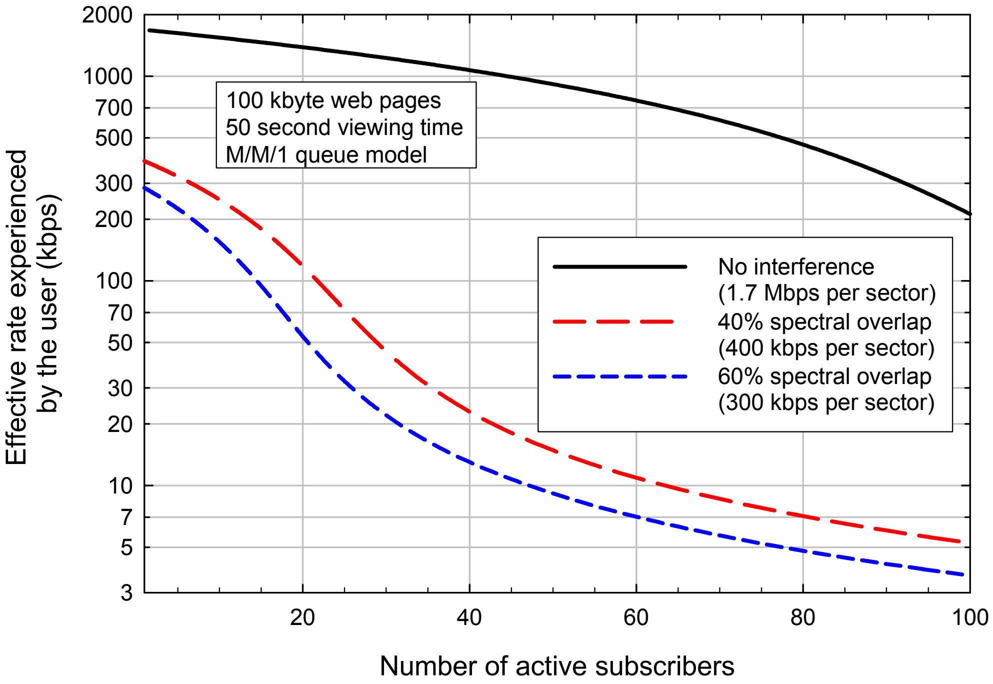


Figure 10: *Comparison for all three rates based on the M/M/1 queue model.*

Summary and Areas for Further Work

The simple models and examples provided here illustrate some of the basic principles that apply to the "user experience" of high speed data. When the load on the data pipe is below its congestion threshold (N_{break}), queue delays are small and each user experiences a rate close to the maximum. Above the threshold, the effective rate experienced by the user falls off rapidly as load increases.

It is also apparent that for a given loading, R_{eff} varies more than the overall available rate R. This can be seen from Figure 3, where except for very light loading, if R is reduced from 2 Mbps to 1 Mbps, R_{eff} decreases by more than a factor of 2.

Clearly, a number of simplifying assumptions have been made here. The data block size, absorption time, and available rate R have been assumed the same for all users. However, allowing these factors to exhibit random variations is unlikely to significantly change the results qualitatively. Indeed, Figure 2 suggests that the difference between exponentially-distributed service time (the M/M/1 queue) and fixed service time (the M/D/1 queue) will be minor, as long as the average service time is kept constant. In fact, the same approach used here can be used to approximate any service time distribution using the Pollaczek-Khinchin (P-K) formula ([4], p. 187):

$$\overline{n}_q = \rho + \frac{\lambda^2 \overline{t_{svc}^2}}{2(1-\rho)} \tag{23}$$

where $\overline{t_{svc}^2}$ is the mean squared service time (the second moment of the service time distribution). For exponentially-distributed service time with a mean of τ, the PDF is $f_{t_{svc}}(t) = e^{-t/\tau}/\tau$ and $\overline{t_{svc}^2} = 2\tau^2$, giving (1). As another example, for service time that is uniformly-distributed between 0 and 2τ, $\overline{t_{svc}^2} = 4\tau^2/3$. Applying this to the P-K formula in (23) gives

$$\overline{n}_q = \frac{\rho - \rho^2/3}{1-\rho} \text{ (uniformly distributed service time).} \tag{24}$$

This is between the results for the fixed and exponentially-distributed service times.

Clearly, there are a number of possible ways in which this simple analysis can be enhanced and extended. More sophisticated traffic models could be used, including variations in the idle (absorption) time, the data block length, and different (non-Poisson) arrival statistics. Also, variations in the rate *R* available to different subscribers (with different SINRs) could be taken into account. The statistics of *R* and of *L* would together determine the service time distribution.

References

[1] "Over Subscription Considerations for DSL Services," Mike Spivey, Qwest, available at https://apps.qwest.com/qhost/content/oversubscription.htm.

[2] www.clickz.com/stats/big_picture/traffic_patterns/article.php/3395351

[3] www.pantos.org/atw/35654.html

[4] Leonard Kleinrock, *Queueing Systems, Volume I: Theory*, New York: Wiley, 1975, ISBN 0-471-49110-1.

[5] Robert B. Cooper, *Introduction to Queueing Theory,* Washington, D.C.: CeePress, third edition, 1990, ISBN 0-941893-03-0.

[6] Ivica Kostanic and Dan McKenna, "Evaluation of the ATG Spectrum Migration Concept," March 10, 2004, AirCell report to the FCC, WT Docket 03-103.

[7] Anthony A. Triolo and Jay E. Padgett, "Coexistence Analysis for Multiple Air-to-Ground Systems," June 3, 2004, Verizon Airfone report to the FCC, WT Docket 03-103.